Pragmatic Reasoning for Irony Detection with Large Language Models in English and Norwegian

Margareta Berg[1], Ildikó Pilán[2], Ingrid Lossius Falkum[1], and Pierre Lison[1,2]
[1]University of Oslo, Oslo, Norway
[2]Norwegian Computing Center, Oslo, Norway
margareta.berg@ifikk.uio.no pilan@nr.no i.l.falkum@ifikk.uio.no plison@nr.no

Abstract

This study investigates the 'pragmatic abilities' of large language models (LLMs) – both standard and reasoning-optimized – across two languages (English and Norwegian). Based on an existing experimental study on children's irony comprehension, we found that LLMs largely identified irony, but performance was poorer in Norwegian due to translation challenges.

1 Introduction

Verbal irony – such as saying "Good job!" to someone who has just failed a task – is among the most complex pragmatic phenomena to master, requiring listeners to infer the speaker's true communicative intent beyond the literal meaning of the utterance. Theories of irony processing in adults suggest that the ironical speaker tacitly echoes a thought (a belief, intention, or norm-based expectation) that they attribute to someone else while simultaneously conveying a dismissive attitude towards that thought (e.g., in the ironical utterance "Good job!", the speaker could be echoing an earlier claim of the addressee of being well-coordinated and never making messes) (Wilson and Sperber, 2012). This complexity is taken to be the main reason why verbal irony comprehension is a relatively late acquisition, emerging around the age of 5 to 6 years and developing further into adolescence[1].

Although a few benchmarks have been developed to assess the capacity of LLMs to process irony and other pragmatic phenomena (Sravanthi et al., 2024; Ma et al., 2025), those are rarely connected to the broader literature and empirical studies in experimental pragmatics. Furthermore, although reasoning-optimized LLMs have emerged as one of the key technological advances in NLP over the past year (Xu et al., 2025), their pragmatic abilities remain underexplored, particularly in a cross-lingual perspective.

The goal of this study is to start filling those gaps. Specifically, we looked at irony detection, and focused on the following questions: (1) Is there a difference when conducting irony-related tasks with LLMs in English and in Norwegian? (2) How do reasoning models compare to their non-reasoning counterparts in irony detection? (3) What similarities and differences emerge between the performance of LLMs and human participants across age groups when responding to the same tasks?

To address these questions, we adapted experimental materials from a prior study on irony in children and adults (Köder and Falkum, 2021) for use with LLMs[2]. Our contributions include preliminary results about how LLMs handle irony-related questions in two different languages and compared to human subjects from different age groups.

2 Related Work

The computational modeling of irony and related pragmatic phenomena is challenging, although models tailored for these have been proposed (Zeng and Li, 2022). Recent NLP work on irony detection has leveraged pretrained transformers like BERT (Devlin et al., 2019), improving performance across languages by incorporating syntactic features (Cignarella et al., 2020), though later studies revealed biases linking irony to strong positive sentiment (Maladry et al., 2023). Hu et al. (2023) compared LLMs and humans and found that LLMs achieve high accuracy, mirror adult human error patterns, and show similar sensitivity to linguistic cues when processing pragmatic phenomena. While multi-modal irony detection with visual data has been explored, textual cues alone were found to often provide sufficient context (Tomás et al., 2023).Yi et al. (2025) showed LLMs with in-context learning can match fine-tuned mod-

[1]See Falkum and Köder (2024) for a review.

[2]The dataset is available at: `https://github.com/IldikoPilan/llm_irony/`

Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue, September 3–5, 2025, Bielefeld, Germany.

els while providing more generalizable, human-like explanations, grounded semantically and affectively. Finally, agent-based frameworks simulating human-like, multi-perspective reasoning have been shown to enhance performance and interpretability in irony detection (Liu et al., 2025).

3 Experimental Setup

Data The material, adapted from a previous pragmatics experiment on irony and perspective-taking in children (Köder and Falkum, 2021), were centered around simple situations involving a child and an adult. In the LLM-adapted version, each task was subdivided into two prompts: one consisting of a short context and a question about the adult speaker's intent, and another containing the child's action, the adult's reaction and a question about the adult's emotion. We added two follow-up questions for each task for investigating the presence of irony with LLMs: an indirect and a direct one, see Table 1 in Appendix A for an example. We complemented the original 12 stories with 24 new unique stories. The final dataset thus comprised 108 items derived from 36 unique stories, each associated with one of three possible outcomes: irony, praise, or criticism – the latter two representing non-ironic reactions. In constrast to the original study with human subjects, which also included images, we employed text-only input, adding text descriptions of the images whenever needed.

Models We compared V3 (Liu et al., 2024) and the reasoning R1 model (Guo et al., 2025) developed by DeepSeek, as well as Gemini 2.5 Flash and the reasoning Pro model (Gemini Team, 2025) from Google. Messages were submitted to the LLM via OpenRouter's[3] API with the full history per item (see Table 1 in the Appendix). We prompted the models to give a structured response of a single word and a short explanation. The token limit for the model reasoning was set to 1000.

4 Results and Discussion

Overall, the four tested models successfully chose the target pragmatic interpretation in most cases, with only 5.1% and 2.9% incorrect answers on average for the indirect and the direct irony question respectively (see Appendix B for detailed results). As the original study was tailored to assess children's understanding of irony and thus contained relatively clear-cut cases, these results are not unexpected. Chi-square tests showed that for the indirect irony question, both LLM type ($p < 0.001$) and language ($p = 0.009$) significantly affected the rate of incorrect responses, with a higher error rate for English than for Norwegian. For the direct irony question, the only factor showing a marginal effect ($p = 0.043$) was the LLM family.

Qualitative analysis We observed that the Gemini models applied to the English data tended to deny that the speaker meant exactly what was said, even in non-ironic cases. The Gemini models identified the utterances as understatements or factual statements used to express more than the literal interpretation. The Gemini models applied to the Norwegian data showed similar results, but with more instances of hallucination. Furthermore, the Gemini models misinterpreted some Norwegian words when translating into English during reasoning, leading to a slight increase in incorrect responses. The impact of reasoning in a language different from that of the user interaction has been shown in previous work (Qi et al., 2025). We tested prompts that explicitly instructed reasoning in Norwegian, but Gemini 2.5 Pro consistently reasoned in English, while R1 did so in Norwegian only 8% of the time. The chat models were slightly better at detecting irony than the reasoning ones. When the chat models failed to identify irony, they seemed to misinterpret the speaker's mental states and their awareness of the context (see Gemini 2.5 Flash in Table 2 in Appendix B). Compared to the original experiment, we observed that the DeepSeek models perform similarly to adults when applied to the English data, while the other models' performances can be compared to 7 or 8 year old children.

5 Conclusions

We presented preliminary results comparing reasoning-based models to their non-reasoning counterparts for detecting irony in English and Norwegian, using material previously applied with adults and children. While all models demonstrated strong irony detection performance, we observed both quantitative and qualitative differences between the two languages. Notably, error patterns revealed challenges such as translation mismatches between the English reasoning process and Norwegian prompts and responses, misinterpretation of irony as understatement, and incorrect assumptions about the parents' understanding of the situation.

[3] https://openrouter.ai/

Limitations

The generalizability of results may be limited by the relatively small number of contexts (and corresponding questions) included in the dataset, along with their similarity, which were all typical adult–child interactions.

References

Alessandra Teresa Cignarella, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, Paolo Rosso, and Farah Benamara. 2020. Multilingual irony detection with dependency syntax and neural models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1346–1358.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Ingrid Lossius Falkum and Franziska Köder. 2024. Investigating irony comprehension in children: Methods, challenges, and ways forward. In *Studying Verbal Irony and Sarcasm: Methodological Perspectives from Communication Studies and Beyond*, pages 145–173. Springer.

Google Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Technical report*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213.

Franziska Köder and Ingrid Lossius Falkum. 2021. Irony and perspective-taking in children: The roles of norm violations and tone of voice. *Frontiers in Psychology*, 12:624604.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ziqi Liu, Ziyang Zhou, and Mingxuan Hu. 2025. CAF-I: A collaborative multi-agent framework for enhanced irony detection with large language models. *arXiv preprint arXiv:2506.08430*.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696.

Aaron Maladry, Els Lefever, Cynthia Van Hee, and Veronique Hoste. 2023. A fine line between irony and sincerity: Identifying bias in transformer models for irony detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 315–324.

Jirui Qi, Shan Chen, Zidi Xiong, Raquel Fernández, Danielle S Bitterman, and Arianna Bisazza. 2025. When models reason in your language: Controlling thinking trace language comes at the cost of accuracy. *arXiv preprint arXiv:2505.22888*.

Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024. PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities. *arXiv [cs.CL]*.

David Tomás, Reynier Ortega-Bueno, Guobiao Zhang, Paolo Rosso, and Rossano Schifanella. 2023. Transformer-based models for multimodal irony detection. *Journal of Ambient Intelligence and Humanized Computing*, 14(6):7399–7410.

Deirdre Wilson and Dan Sperber. 2012. Explaining irony. *Meaning and relevance*, pages 123–145.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, and 1 others. 2025. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

Peiling Yi, Yuhan Xia, and Yunfei Long. 2025. Irony detection, reasoning and understanding in zero-shot learning. *IEEE Transactions on Artificial Intelligence*.

Qingcheng Zeng and An-Ran Li. 2022. A survey in automatic irony processing: Linguistic, cognitive, and multi-X perspectives. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 824–836, Gyeongju, Republic of Korea.

A Input Example

Table 1 shows a full input subdivided into its constituting messages, submitted one by one to the LLM (introduction and intention messages were submitted at the same time). The prompt included also instructions about the desired output format (a JSON object) and LLM-generated responses to any previous questions for the same item.

Type	Message
Introduction	You will soon hear a short story. You will be asked to answer some questions during the story, so it is important that you pay close attention.
Intention	Heidi is celebrating her birthday and wants to invite her friends. Mum says to Heidi: "Here are the invitations for your birthday. Remember to hand them out at school today." Heidi says: "Sure, I'll remember to hand out the invitations." What does Heidi's mum want? A: Mum wants Heidi to hand out the invitations at school. B: Mum doesn't want Heidi to hand out the invitations at school.
Emotion	Heidi forgets to hand out the invitations, and brings them back home in her backpack. Mum says to Heidi: "That's great. You remembered to hand out the invitations." How is mum feeling inside? Is she happy or angry?
Irony (indirect)	Did the grown-up mean exactly what they said?
Irony (direct)	Is there any verbal irony in the adult's reaction?

Table 1: Input example employed for the experiments, segmented into user messages.

B Detailed Results

Quantitative results per language, LLM family (Deepseek or Gemini) and model type (chat or reasoning-optimized) are presented in Figure 1. For DeepSeek models, we used the free versions of R1 0528 and V3 0324.

Finally, Table 2 presents a particular error pattern in which one of the non-reasoning models (Gemini 2.5 Flash) seemed to misinterpret the speaker's mental states and their understanding of the context.

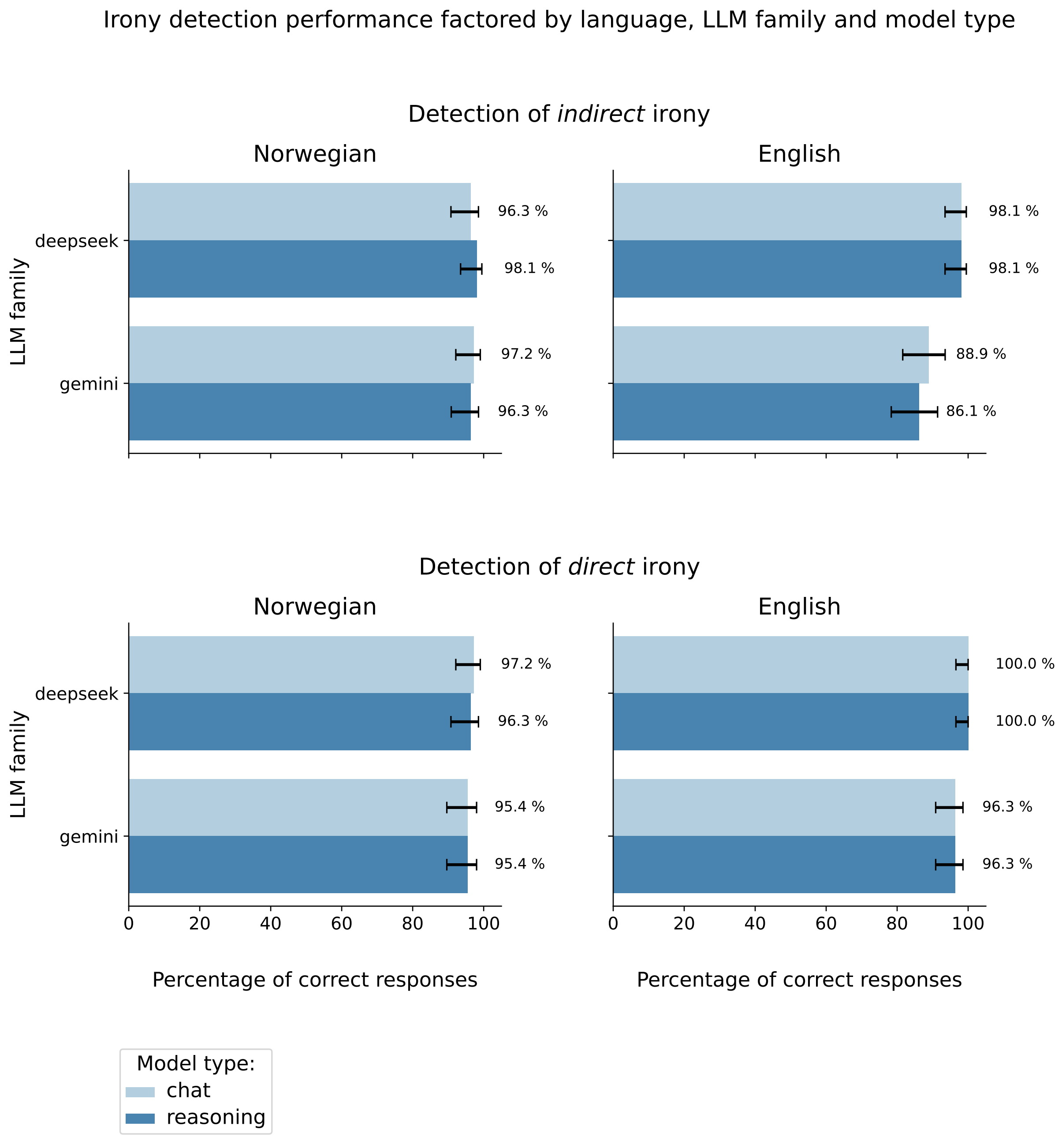

Figure 1: Percentage of correct responses on irony detection per LLM (Gemini or Deepseek) and model type (chat or reasoning-optimized). The percentages are provided both for the indirect questions on the presence of irony (top) as well as the direct questions (bottom), and for Norwegian (left) versus English (right) . The results are provided with 95 % confidence intervals estimated using the Wilson score method.

Model	Resp.	Explanation	Reasoning
DeepSeek V3	No	Mum's statement 'That's great' was sarcastic, as Heidi did not actually hand out the invitations, so she did not mean it literally.	-
Gemini 2.5 Flash	Yes	The grown-up meant exactly what they said based on her understanding of the situation. She genuinely believed Heidi had handed out the invitations and was expressing her happiness about it.	-
DeepSeek R1	No	Mum used sarcasm to express her displeasure, as Heidi forgot to hand out the invitations despite agreeing to do so. Her words contradict her true feelings.	Alright, so the user is asking if the grown-up (Heidi's mum) meant exactly what she said. Let's go back to the story. In the first part, Mum told Heidi to hand out the invitations at school, and Heidi agreed but then forgot and brought them back home. When Mum says, "That's great. You remembered to hand out the invitations," it's pretty clear Heidi didn't do it. So Mum's statement is sarcastic here. The previous answer correctly identified that Mum was feeling angry because of the sarcasm. Now, the user is asking if Mum meant her words literally. Since sarcasm involves saying the opposite of what you mean to highlight the opposite situation, the answer would be 'No'. Mum didn't truly think it was great; she was upset Heidi forgot. So the reply should state that Mum didn't mean her words literally and explain the sarcasm.
Gemini 2.5 Pro	No	Mum was using sarcasm. She said the opposite of what was true to show she was disappointed or angry that Heidi had forgotten the invitations.	**Deciphering the Question** I'm now considering the user's question: "Did the grown-up mean exactly what they said?". It's a bit complex, I think I need to clarify the grown-up's words in the context of the prior snippet. My output must be a JSON object with two keys, which is a constraint I will need to consider when interpreting the question. **Formulating JSON Response** [...] **Confirming Final Structure** I've reviewed the entire process one last time. I'm satisfied with the JSON object: the 'choice' key is "No," and the 'explanation' clearly states that the Mum's words were sarcastic. The final JSON structure now perfectly aligns with the initial requirements.

Table 2: Example of responses to the indirect irony question 'Did the grown-up mean exactly what they said?' for the experimental item containing irony from Table 1. Gemini 2.5 Flash provides an incorrect response.